Received: 4 April 2019 | Revised: 13 June 2019 | Accepted: 6 July 2019

DOI: 10.1111/jopy.12500

ORIGINAL ARTICLE

WILEY

Global sex differences in personality: Replication with an open online dataset

Tim Kaiser[1] | **Marco Del Giudice**[2] | **Tom Booth**[3]

[1]Department of Psychology, University of Salzburg, Salzburg, Austria

[2]Department of Psychology, University of New Mexico, Albuquerque, New Mexico

[3]Department of Psychology, University of Edinburgh, Edinburgh, UK

Correspondence
Tim Kaiser, Department of Psychology, University of Salzburg, Hellbrunnerstrasse 34, 5020 Salzburg, Austria.
Email: tim.kaiser@sbg.ac.at

Abstract

Objective: Sex differences in personality are a matter of continuing debate. In a study on the United States standardization sample of Cattell's 16PF (fifth edition), Del Giudice and colleagues (2012; *PLoS ONE, 7,* e29265) estimated global sex differences in personality with multigroup covariance and mean structure analysis. The study found a surprisingly large multivariate effect, $D = 2.71$. Here we replicated the original analysis with an open online dataset employing an equivalent version of the 16PF.

Method: We closely replicated the original MG-MCSA analysis on $N = 21{,}567$ U.S. participants (63% females, age 16–90); for robustness, we also analyzed $N = 31{,}637$ participants across English-speaking countries (61% females, age 16–90).

Results: The size of global sex differences was $D = 2.06$ in the United States and $D = 2.10$ across English-speaking countries. Parcel-allocation variability analysis showed that results were robust to changes in parceling (U.S.: median $D = 2.09$, IQR [1.89, 2.37]; English-speaking countries: median $D = 2.17$, IQR [1.98, 2.47]).

Conclusions: Our results corroborate the original study (with a comparable if somewhat smaller effect size) and provide new information on the impact of parcel allocation. We discuss the implications of these and similar findings for the psychology of sex differences.

KEYWORDS

effect size, gender differences, Mahalanobis' *D*, multivariate, sex differences

1 | INTRODUCTION

More than a century after the publication of the first systematic studies (see Allen, 1930; Thompson, 1903), sex differences in personality continue to be passionately debated. Taken as groups, do men and women show substantially different patterns of feelings, thoughts, and behaviors? Or does the overlap between the sexes dwarf whatever discrepancies exist? The latter is closer to the prevailing view in psychology, which embraces the "gender similarities hypothesis" that men and women are similar on most psychological variables (Hyde, 2005, 2014). Personality has implications for a multitude of important life outcomes, from physical and mental health to occupational choices and work performance (Friedman & Kern, 2014; Kotov, Gamez, Schmidt, & Watson, 2010; Soto, 2019). Moreover, current ideas about the causes and effects of gender stereotypes are inevitably shaped by assumptions about the magnitude of sex differences across domains (e.g., Fiske, 2017; Haines, Deaux, & Lofaro, 2016; Hyde, Bigler, Joel, Tate,

& van Anders, 2019; Zell, Strickhouser, Lane, & Teeter, 2016). For all these reasons, quantifying sex differences as accurately and meaningfully as possible is a crucial task for personality psychology.

Studies conducted with the Five-Factor Model of personality (also known as the Big Five) show that, averaging across countries, women score about 0.4–0.5 standard deviations higher in Agreeableness and Neuroticism (Cohen's $d \approx -0.40$ to -0.50; by convention, we use negative values to indicate higher scores in females). Sex differences in the other domains—Conscientiousness, Extraversion, and Openness—are smaller, typically 0.1 *SD* or less (see Del Giudice, 2015; Hyde, 2014; Kajonius & Mac Giolla, 2017; Lippa, 2010; Löckenhoff et al., 2014; Schmitt, Realo, Voracek, & Allik, 2008). Assuming normal distributions, these figures imply that the overlap between male and female distributions ranges from about 80% to almost 100%, depending on the trait. Based on these findings, some researchers have declared that the evidence overwhelmingly indicates similarity, and refutes the view that human behavior is sexually dimorphic to any significant degree (e.g., Hyde, 2014; Hyde et al., 2019).

From a methodological standpoint, the standard approach to measuring sex differences suffers from three important limitations. First, the effect sizes in the literature are typically calculated directly from observed scores, with no correction for measurement error (e.g., Hyde, 2005, 2014; Zell, Krizan, & Teeter, 2015). This can lead to underestimate their magnitude by a substantial margin. To correct for measurement error, one can disattenuate the observed effect sizes with reliability coefficients (e.g., Cronbach's α), or estimate the effects from latent, error-free variables (see Del Giudice, 2019). Second, focusing on broad personality factors like the Big Five domains misses much of the structure of sex differences in personality, which become more apparent at the level of narrower traits (*aspects* or *facets* in the Big Five model). Moreover, it is sometimes the case that different facets of the same domain (e.g., intellect/ideas vs. aesthetics/feelings in the Openness domain) show sex differences in opposite directions, which tend to cancel each other out at the level of broad factors (Costa, Terracciano, & McCrae, 2001; Del Giudice, 2015; Kajonius & Johnson, 2018; Soto, John, Gosling, & Potter, 2011; Weisberg, DeYoung, & Hirsh, 2011). Third, in the standard approach individual personality traits are considered one at a time, or simply averaged together (e.g., Zell et al., 2015). However, differences across multiple correlated traits can add up to yield a much larger effect size in the multivariate space. This problem can be readily solved by calculating a multivariate effect size. The natural choice for sex differences is Mahalanobis' *D*, which generalizes Cohen's *d* for two or more correlated variables. *D* is the unsigned standardized distance between the centroids (multivariate means) of the two groups, and has the same basic interpretation as *d* (see Del Giudice, 2009, 2013, 2019; Hess, Hogarty, Ferron, & Kromrey, 2007; Olejnik & Algina, 2000).

To address these limitations, Del Giudice and colleagues (Del Giudice, Booth, & Irwing, 2012) used *D* to estimate the size of global (i.e., multivariate) sex differences in the United States standardization sample of the fifth edition of Cattell's Sixteen Personality Factors questionnaire (16PF), a demographically representative sample with $N = 10{,}261$ (50.1% females; age 16–90 years). The 16PF comprises 15 narrow personality factors, allowing a more fine-grained analysis of sex differences than the Big Five domains (the remaining factor in the 16PF is a measure of cognitive ability). Multigroup covariance and mean structure analysis (MG-CMSA) was used to estimate latent differences and correlations, and test for measurement invariance and equality of correlation matrices. Both the measurement model and the correlation matrix were invariant across sexes. The multivariate difference was surprisingly large, amounting to $D = 2.71$. This effect size implies a marked statistical separation between males and females: the estimated proportion of overlap is about 18% of each distribution, and about 10% of the joint distribution (assuming multivariate normality; for details see Del Giudice, 2019).

The study by Del Giudice and colleagues (2012) was the first to challenge the idea that sex differences in personality are small to moderate in magnitude. More recently, researchers have started to apply the concept of global sex differences to datasets based on the Big Five model. In a cross-cultural study, Mac Giolla and Kajonius (2018) computed *D* from observed scores on 30 facets of the Big Five, without error correction. The size of global sex differences in the United States was $D = 1.25$, similar to the uncorrected effect ($D = 1.49$) in Del Giudice and colleagues (2012). Kaiser (2019) also considered 30 Big Five facets, but used latent scores estimated with MG-CMSA and found $D = 2.16$ in the United States. These results corroborate the initial findings by Del Giudice and colleagues (2012), and are consistent with the idea that sex differences in personality are much larger than previously assumed. However, the original study has yet to be replicated using the 16PF model.

In the present study, we set out to replicate the analysis by Del Giudice and colleagues (2012) with a large dataset available from the Open Source Psychometrics Project (https://openpsychometrics.org). The dataset employs an equivalent version of Cattell's 16PF constructed with items from the International Personality Item Pool (IPIP; Goldberg, 1999), and comprises a total of 49,159 responses from various countries (more details in the Methods section). In accord with the original study, we focused primarily on the United States; we included other English-speaking countries in a secondary analysis. After replicating the original study as closely as possible, we investigated the robustness of our findings to changes in the allocation of items to parcels (parcel-allocation

variability [PAV]; Sterba & MacCallum, 2010) by randomly allocating items to parcels and computing the resulting distribution of *D* values. Our working hypothesis was that the size of global sex differences in personality and the pattern of univariate sex differences across individual traits would be similar to those of the original study.

2 | METHODS

2.1 | Samples and measures

2.1.1 | Sample selection and data cleaning

The 16PF dataset was retrieved in December 2018 from the Open Source Psychometrics website. The full dataset ($N = 49{,}159$) was uploaded to the website on May 14, 2014. The answers were given by anonymous users who took the free questionnaire on the website. At the beginning of the questionnaires, respondents were informed that their answers would be used for research purposes, as follows: "your answers on this test will be stored and used for research, and possibly shared in a way that preserves your anonymity."

We began by selecting two samples from the dataset: (1) respondents from the United States ($N = 23{,}701$; 64% females), and (2) respondents from English-speaking countries ($N = 34{,}625$; 62% females): United States, United Kingdom ($N = 4{,}630$), Canada ($N = 2{,}420$), Australia ($N = 2{,}280$), Ireland ($N = 372$), New Zealand ($N = 287$), and others ($N = 83$). Geographical location had been determined based on the IP address of the connection (note that the dataset does *not* include IP addresses, but only country codes). Because the main goal of the present study was replication, we did not consider data from non-English-speaking countries. In addition, the questionnaire was only administered in English, and there are reasonable concerns about the validity of online responses by participants whose primary language is not English (Feitosa, Joseph, & Newman, 2015). Respondents who did not indicate their sex were excluded at this stage. To replicate the original analysis (Del Giudice et al., 2012), we excluded respondents under 16 and over 90 years of age.

To address the problem of careless responding, we excluded respondents with 25 or more missing answers, and those who responded to all the items with the same end-of-scale answer (either 1 or 5). The questionnaire also included a follow-up question asking respondents to rate the accuracy of their answers on a 0%–100% scale. We only retained participants who rated their accuracy at 50% or more. Together, these criteria led to the exclusion of approximately 9% of the initial cases. Since the exclusion criteria we applied were not pre-registered, we also repeated the analyses on the full U.S. and English-speaking samples as a robustness check.

The size of the final U.S. sample was $N = 21{,}567$, with 63% females. The mean age was 26.6 years in males ($SD = 12.2$) and 25.5 years in females ($SD = 11.3$). The size of the final English-speaking sample was $N = 31{,}637$, with 61% females. The mean age was 26.7 years in males ($SD = 12.1$) and 25.9 years in females ($SD = 11.5$). These samples were used to perform the main analyses described below; the analyses were then repeated on the full (unselected) samples. All analyses were performed in R 3.5.2 (R Core Team, 2018). To ensure reproducibility, both the original data and the annotated R scripts of our analyses are available at https://osf.io/4sq79/?view_only=51e0ba73a0234840a2902f586e52f4c2 (anonymized link).

2.1.2 | Personality measures

The questionnaire comprises 163 items organized into an Intellect scale (Factor B or Reasoning in the original 16PF) and 15 primary personality scales with 10 items each. For each scale, we report the original name in parentheses. Warmth (A, Warmth), Emotional Stability (C, Emotional Stability), Assertiveness (E, Dominance), Gregariousness (F, Liveliness), Dutifulness (G, Rule-Consciousness), Friendliness (H, Social Boldness), Sensitivity (I, Sensitivity), Distrust (L, Vigilance), Imagination (M, Abstractness), Reserve (N, Privateness), Anxiety (O, Apprehension), Complexity (Q1, Openness to change), Introversion (Q2, Self-Reliance), Orderliness (Q3, Perfectionism), and Emotionality (Q4, Tension). The items were selected from the IPIP pool to match the content of the original 16PF scales. The response format is on a 5-point Likert scale, from "strongly disagree" to "strongly agree." In the original scale construction study, correlations between the original and equivalent personality scores ranged from .75 to .99 (corrected for unreliability; Goldberg, 1999). Cronbach's α values for observed scores in the present study are reported in Tables 1 and 4.

2.2 | Data analysis

2.2.1 | Model fitting and invariance tests

For each of the two samples (U.S. and English-speaking), we fit a set of MG-CMSA models in which 3 item parcels (see below) loaded on each of 15 correlated factors, following the a priori structure of the 16PF. No cross-loadings or correlated errors were modeled. Models were fit with package *lavaan* (v.0.6-3; Rosseel, 2012) using a maximum likelihood estimator with robust standard errors (MLR). Goodness of fit was evaluated with χ^2, CFI, NNFI, RMSEA, and SRMR. There is much debate on the use of, and exact values of, cut-offs for model fit. Here we applied the commonly suggested criteria of > 0.90–0.95 indicating acceptable to excellent fit for the CFI and NNFI, and 0.06–0.08 indicating acceptable to

TABLE 1 Correlations, univariate effect sizes, and internal consistency values for observed scores (U.S. sample)

	A.	C.	E.	F.	G.	H.	I.	L.	M.	N.	O.	Q1.	Q2.	Q3.	Q4.
A. Warmth		.23	.15	.37	−.23	.49	.26	−.42	−.04	−.46	−.02	−.26	−.39	−.08	−.39
C. Emot. Stability	.25		.34	.17	−.25	.43	−.01	−.45	−.22	−.27	−.74	−.11	−.21	−.17	−.54
E. Assertiveness	.19	.40		.29	.08	.50	.03	.00	.01	−.29	−.38	−.24	−.11	−.25	.03
F. Gregariousness	.44	.26	.34		.13	.63	.00	−.17	.19	−.39	−.12	−.17	−.50	.11	−.08
G. Dutifulness	−.22	−.25	.05	.09		−.08	.02	.33	.50	.11	.03	−.26	.15	.34	.27
H. Friendliness	.53	.46	.54	.67	−.13		.04	−.35	−.10	−.62	−.34	−.17	−.54	−.10	−.24
I. Sensitivity	.28	−.05	.03	−.01	.08	.07		−.14	.23	−.15	.05	−.47	.11	.01	−.14
L. Distrust	−.46	−.42	−.03	−.24	.36	−.39	−.12		.22	.42	.30	.07	.35	.02	.53
M. Imagination	−.04	−.26	−.01	.11	.56	−.15	.25	.28		.10	.12	−.46	.20	.34	.10
N. Reserve	−.54	−.29	−.33	−.46	.16	−.66	−.16	.45	.13		.14	.13	.47	.04	.16
O. Anxiety	−.07	−.76	−.42	−.20	.09	−.40	.04	.29	.21	.18		.14	.09	.06	.46
Q1. Complexity	−.31	−.13	−.27	−.18	−.25	−.21	−.53	.09	−.42	.20	.13		−.07	−.08	.24
Q2. Introversion	−.46	−.26	−.19	−.54	.21	−.60	.09	.44	.26	.53	.16	−.03		−.03	.19
Q3. Orderliness	−.08	−.20	−.23	.12	.41	−.11	.02	.07	.37	.07	.10	−.07	.02		.03
Q4. Emotionality	−.43	−.53	−.02	−.15	.27	−.29	−.21	.56	.16	.25	.48	.24	.25	.06	
d	−0.35	+0.31	+0.19	+0.01	+0.32	0.00	−0.80	+0.05	+0.11	+0.15	−0.51	−0.20	−0.01	+0.09	−0.09
α	.84	.87	.84	.80	.84	.91	.68	.87	.80	.88	.85	.79	.85	.80	.81

Note: Females above the diagonal; males below the diagonal. Positive values of *d* indicate that males score higher than females. Correlations and effect sizes are uncorrected (i.e., not adjusted for score unreliability). The letters associated with each scale are the conventional trait identifiers in the 16PF model.

excellent fit for the RMSEA and SRMR (e.g., Hu & Bentler, 1999; Schermelleh-Engel, Moosbrugger, & Müller, 2003).

Measurement invariance across sexes was tested by fitting three models: one without constraints (configural invariance), one with equal loadings (metric invariance), and one with equal loadings and intercepts (scalar invariance). In deciding whether invariance held across models, we inspected the change in model fit. Simulation studies have suggested that a change if CFI of −0.01 or less, and changes in RMSEA of less than or equal to 0.015 are suggestive that invariance holds (Chen, 2007).

The invariance of interfactor correlation matrices across sexes was tested by adding an equality constraint to the model and comparing fit indices; in addition, we calculated Tucker's congruence coefficient (*CC*) between the model-estimated male and female matrices to quantify their similarity (see Del Giudice, 2019).

2.2.2 | Parceling

To replicate the original study by Del Giudice et al. (2012) as closely as possible, in the main analysis we employed the Single Factor method (Landis, Beal, & Tesluk, 2000) to create three parcels for each personality scale. Parcels reduce the number of parameters to estimate and often show improved characteristics compared with individual items (e.g., higher reliability, lower likelihood of distributional violations; see Little, Rhemtulla, Gibson, & Schoemann, 2013). However, they also introduce an additional source of variation, as estimated model parameters may vary across different potential allocations of items to parcels. The impact of PAV has been shown to be stronger when sample size is small, item communalities are low, or there are small numbers of parcels and/or items per parcel (Sterba & MacCallum, 2010; see also Sterba, 2019). To assess the robustness of our main findings with respect to PAV, we generated 100 random item allocations with three parcels per factor using package *semTools* v. 0.5-1 (Jorgensen, Pornprasertmanit, Schoemann, & Rosseel, 2018). For each allocation, we fit a set of MG-CMSA models (see above), estimated parameters from the scalar-invariant model, and examined the resulting distribution of effect sizes.

2.2.3 | Effect sizes and related statistics

To compute effect sizes and related statistics, we employed the R scripts available at https://doi.org/10.6084/m9.figshare.7934942.v1 and described in Del Giudice (2019). For latent variable models, we computed Mahalanobis' *D* with exact confidence intervals (Reiser, 2001) from group differences and interfactor correlations estimated from the scalar-invariant models. For observed scores, we obtained bootstrap confidence intervals from 10,000 samples (Kelley, 2005). Values of *D* were then used to estimate the overlapping coefficient *OVL* (the proportion of each distribution shared with the other), Cohen's coefficient of overlap OVL_2 (the shared proportion of the joint distribution, $1–U_1$ in Cohen, 1988), and the common language effect size *CL* (in this case, the probability that a randomly picked male will show a more male-typical profile than a randomly picked female, and vice versa; see Del Giudice, 2019; McGraw & Wong, 1992). Finally, we calculated coefficients H_2 and EPV_2 to quantify heterogeneity in the contribution of individual personality traits to global sex differences (as measured by *D*). Coefficient H_2 ranges from 0 (maximum homogeneity; all variables contribute equally) to 1 (maximum heterogeneity; the totality of the effect is explained by just one variable). The "equivalent proportion of variables" coefficient EPV_2 (also on a 0–1 scale) estimates the proportion of equally contributing variables that would produce the same amount of heterogeneity, if the other variables in the set made no contribution. For example, $EPV_2 = .30$ means that the same amount of heterogeneity would obtain if 30% of the variables contributed equally to the overall effect and the remaining 70% made no contribution (Del Giudice, 2017, 2018, 2019).

3 | RESULTS

3.1 | United States

3.1.1 | Main analysis

We started our analysis of sex differences from observed scores on the fifteen 16PF scales. Correlations and univariate standardized differences are reported in Table 1. The pattern of observed univariate effect sizes in this sample was very similar to that in the original study by Del Giudice et al. (2012), with a correlation of .95 across personality factors. Male and female correlation matrices showed high similarity ($CC = .99$). The uncorrected size of global sex differences was $D = 1.18$, with 95% CI [1.143, 1.207]. Correction for unreliability raised the effect size to $D = 1.68$. The corresponding overlapping coefficients for uncorrected scores were $OVL = .56$ and $OVL_2 = .39$; for corrected scores they were $OVL = .40$ and $OVL_2 = .25$. In the common language effect size metric, these values translate to $CL = .80$ for uncorrected scores and .88 for corrected scores.

Fit statistics for the main set of MG-CMSA models in the U.S. sample are reported in Table 2. The baseline configural model showed acceptable to excellent fit according to all indices except the NNFI, which fell just below the .90 cut-off. Model fit for the sequentially constrained models suggested that invariance held at all levels. Scalar invariance was met (Model 3 in Table 2), and the resulting mean and correlation estimates (Table 3) were used to compute effect sizes. The correlation between latent univariate effect sizes in this sample and in the original study was .90. Estimated

TABLE 2 Goodness-of-fit statistics for MG-CMSA models (U.S. sample)

Model	χ^2	*df*	CFI	NNFI	RMSEA	SRMR
1. Configural invariance	56,938.27	1,680	.908	.892	.055	.056
2. Metric invariance	57,633.98	1,725	.907	.893	.055	.059
Δ 1 versus 2			*−.001*	*.001*	*.000*	*.003*
3. Scalar invariance	61,786.38	1,755	.900	.887	.056	.059
Δ 2 versus 3			*−.007*	*−.007*	*.001*	*.000*
4. Equality of covariances	62,448.00	1,860	.899	.893	.055	.061
Δ 3 versus 4			*−.001*	*.006*	*−.001*	*.002*

correlation matrices in males and females were highly similar ($CC = .99$); adding a covariance equality constraint to the model did not appreciably change the goodness-of-fit (Model 4 in Table 2). The size of global sex differences estimated from latent scores was $D = 2.06$, with 95% CI [2.03, 2.10]. Assuming multivariate normality, this corresponds to $OVL = .30$, $OVL_2 = .18$, and $CL = .93$. Heterogeneity coefficients for D were $H_2 = .81$ and $EPV_2 = .24$.

3.1.2 | Robustness checks

The first robustness check we ran was to repeat the MG-CMSA analysis on the full U.S. sample, with no exclusion criteria. The size of sex differences did not change appreciably: uncorrected $D = 1.15$; unreliability corrected $D = 1.66$; CFA-estimated $D = 2.04$.

To assess the impact of PAV on the size of sex differences, we computed D values from 100 models with randomly generated parcels. The median effect size was $D = 2.09$, very close to the one obtained in the main analysis. The interquartile range of D was 1.89–2.37; the full distribution is shown in Figure 1a.

3.2 | English-speaking countries

3.2.1 | Main analysis

Correlations and univariate standardized differences for observed scores are reported in Table 4. Again, male and female correlation matrices showed substantial similarity ($CC = .99$). The uncorrected size of global sex differences was $D = 1.19$, with 95% CI [1.16, 1.21]. Correction for unreliability raised the effect size to $D = 1.69$. The corresponding overlapping coefficients for uncorrected scores were $OVL = .55$ and $OVL_2 = .38$; for corrected scores they were $OVL = .40$ and $OVL_2 = .25$. These values translate to $CL = .80$ for uncorrected scores and .88 for corrected scores.

Fit statistics for the main set of MG-CMSA models in the English-speaking sample are reported in Table 5. The pattern of model fit was identical to the U.S. sample, with an NNFI slightly below the .90 cut-off, and all differences in fit within the criteria suggesting invariance held across sex. Scalar invariance was met (Model 3 in Table 5), and the resulting mean and correlation estimates (Table 6) were used to compute effect sizes. Estimated correlation matrices in males and females were highly similar ($CC = .99$); adding a covariance equality constraint to the model did not appreciably change the goodness-of-fit (Model 4 in Table 5). The size of global sex differences estimated from latent scores was $D = 2.10$, with 95% CI [2.07, 2.13]. Assuming multivariate normality, this corresponds to $OVL = .29$, $OVL_2 = .17$, and $CL = .93$. Heterogeneity coefficients for D were $H_2 = .83$ and $EPV_2 = .23$.

3.2.2 | Robustness Checks

Sex differences in the full English-speaking sample (no exclusion criteria) were almost identical to those in the selected sample: uncorrected $D = 1.17$; unreliability corrected $D = 1.68$; CFA-estimated $D = 2.10$. The median effect size in the PAV analysis was $D = 2.17$, again very close to the value obtained in the main analysis. The interquartile range of D was 1.98–2.47; the full distribution is shown in Figure 1b.

4 | DISCUSSION

In this paper, we sought to replicate the findings by Del Giudice and colleagues (2012) with an open online dataset that employed an equivalent version of the 16PF questionnaire based on IPIP items (Goldberg, 1999). Invariance tests and indices of matrix similarity indicated that the correlational structure of personality was equivalent in the two sexes. This allowed us to aggregate sex differences across fifteen personality traits into a multivariate effect size. In the sample of U.S. participants, we estimated a global sex difference of $D = 2.06$. Assuming multivariate normality, the overlap between the sexes implied by this effect size is about 30% of each distribution and 18% of the joint distribution. Sex differences were similar in the larger

TABLE 3 Correlations and univariate effect sizes for MG-CMSA latent scores (U.S. sample)

	A.	C.	E.	F.	G.	H.	I.	L.	M.	N.	O.	Q1.	Q2.	Q3.	Q4.
A. Warmth		.36	.21	.53	−.31	.66	.37	−.58	−.07	−.59	−.11	−.36	−.53	−.11	−.55
C. Emot. Stability	.36		.47	.22	−.30	.54	−.01	−.58	−.29	−.33	−.90	−.17	−.28	−.26	−.69
E. Assertiveness	.27	.55		.38	.11	.60	.04	−.03	.01	−.34	−.52	−.32	−.15	−.31	.01
F. Gregariousness	.61	.32	.44		.16	.77	−.03	−.24	.21	−.46	−.17	−.21	−.68	.16	−.12
G. Dutifulness	−.27	−.28	.08	.12		−.12	.02	.42	.66	.18	.04	−.32	.18	.47	.33
H. Friendliness	.70	.57	.64	.81	−.15		.04	−.47	−.15	−.70	−.44	−.20	−.68	−.14	−.33
I. Sensitivity	.37	−.05	.04	−.04	.09	.05		−.20	.37	−.20	.07	−.76	.18	.04	−.23
L. Distrust	−.62	−.51	−.04	−.34	.44	−.49	−.16		.30	.53	.39	.12	.46	.06	.68
M. Imagination	−.08	−.34	−.04	.11	.71	−.23	.37	.36		.14	.17	−.60	.26	.45	.15
N. Reserve	−.68	−.34	−.38	−.57	.19	−.75	−.20	.55	.17		.18	.16	.56	.06	.22
O. Anxiety	−.14	−.92	−.57	−.25	.10	−.48	.04	.36	.28	.22		.19	.14	.14	.58
Q1. Complexity	−.42	−.18	−.35	−.23	−.30	−.23	−.79	.13	−.55	.23	.17		−.08	−.11	.32
Q2. Introversion	−.61	−.32	−.23	−.70	.24	−.74	.15	.54	.34	.64	.20	−.05		−.03	.25
Q3. Orderliness	−.12	−.30	−.31	.16	.54	−.17	.03	.12	.51	.10	.19	−.09	.04		.08
Q4. Emotionality	−.57	−.66	−.06	−.20	.33	−.37	−.30	.71	.23	.32	.59	.32	.32	.12	
d	−0.37	+0.32	+0.26	+0.02	+0.37	0.00	−0.92	+0.02	+0.11	+0.16	−0.55	−0.24	−0.01	+0.07	−0.09

Note: Females above the diagonal; males below the diagonal. Positive values of *d* indicate that males score higher than females. The letters associated with each factor are the conventional trait identifiers in the 16PF model.

TABLE 4 Correlations, univariate effect sizes, and internal consistency values for observed scores (English-speaking sample)

	A.	C.	E.	F.	G.	H.	I.	L.	M.	N.	O.	Q1.	Q2.	Q3.	Q4.
A. Warmth		.24	.15	.37	−.22	.49	.26	−.42	−.04	−.46	−.02	−.27	−.39	−.08	−.39
C. Emot. Stability	.25		.36	.18	−.25	.44	−.01	−.46	−.22	−.28	−.74	−.12	−.21	−.17	−.55
E. Assertiveness	.20	.40		.28	.08	.50	.03	−.02	.00	−.30	−.39	−.25	−.10	−.24	.02
F. Gregariousness	.42	.24	.33		.13	.63	−.02	−.16	.18	−.38	−.13	−.16	−.50	.13	−.07
G. Dutifulness	−.23	−.26	.05	.08		−.07	.01	.32	.49	.11	.03	−.25	.14	.34	.26
H. Friendliness	.53	.46	.54	.66	−.13		.03	−.36	−.11	−.62	−.35	−.17	−.54	−.09	−.24
I. Sensitivity	.26	−.05	.04	−.05	.08	.04		−.13	.23	−.14	.05	−.48	.12	.00	−.14
L. Distrust	−.46	−.42	−.03	−.22	.36	−.38	−.10		.23	.42	.30	.07	.35	.02	.54
M. Imagination	−.07	−.29	−.02	.09	.56	−.18	.25	.30		.11	.12	−.45	.22	.33	.11
N. Reserve	−.53	−.29	−.33	−.44	.16	−.66	−.13	.44	.15		.15	.13	.47	.04	.17
O. Anxiety	−.07	−.76	−.42	−.19	.10	−.40	.05	.30	.23	.18		.15	.10	.05	.46
Q1. Complexity	−.29	−.11	−.28	−.15	−.25	−.19	−.56	.07	−.42	.18	.11		−.07	−.08	.24
Q2. Introversion	−.45	−.26	−.19	−.54	.21	−.60	.13	.43	.27	.52	.17	−.05		−.04	.19
Q3. Orderliness	−.10	−.20	−.23	.13	.40	−.11	−.01	.07	.36	.07	.11	−.05	.00		.03
Q4. Emotionality	−.41	−.54	−.02	−.10	.28	−.27	−.20	.56	.19	.23	.48	.22	.23	.08	
d	−0.36	+0.33	+0.22	+0.03	+0.29	+0.03	−0.83	+0.03	+0.06	+0.13	−0.52	−0.17	−0.04	+0.07	−0.09
α	.84	.87	.84	.80	.83	.91	.69	.87	.80	.88	.86	.78	.85	.80	.82

Note: Females above the diagonal; males below the diagonal. Positive values of *d* indicate that males score higher than females. Correlations and effect sizes are uncorrected (i.e., not adjusted for score unreliability). The letters associated with each scale are the conventional trait identifiers in the 16PF model.

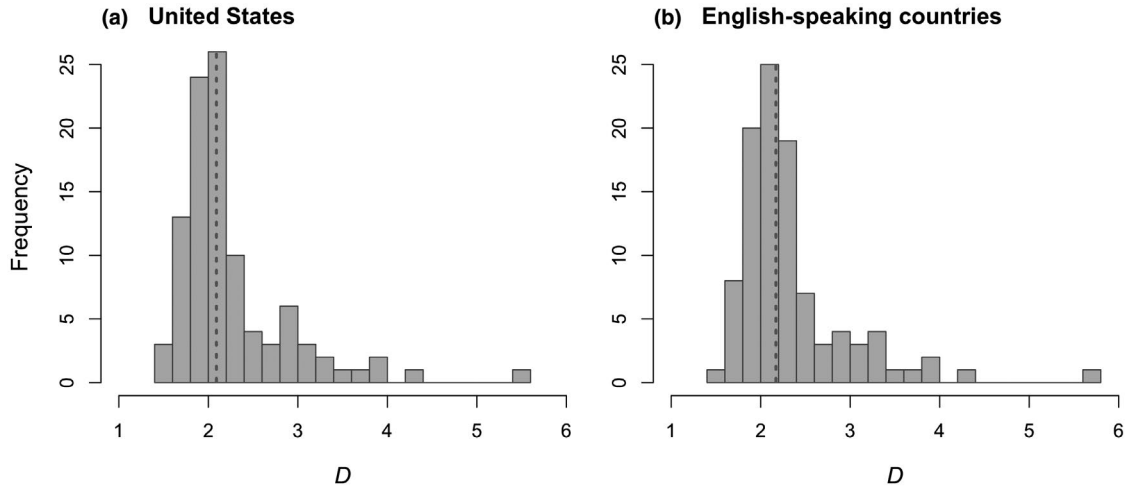

FIGURE 1 Parcel-allocation variability (PAV) analysis of global sex differences. Each panel depicts 100 effect sizes (Mahalanobis' *D*), estimated from MG-CMSA models in which items were randomly allocated to parcels within each personality trait. Median values are shown as dotted lines

sample of respondents from English-speaking countries ($D = 2.10$); both effects were robust to exclusion/inclusion of respondents (based on age and response quality) and to PAV (Figure 1). Predictably, computing effect sizes from estimated means and correlations substantially increased the magnitude of sex differences compared with observed scores.

While substantial, the multivariate effect size estimated in the U.S. sample was 24% smaller than the one in the original study ($D = 2.71$). This difference could be explained by a number of factors. To begin, the two questionnaires were not identical. The coverage of specific personality factors may differ somewhat between the original 16PF and the IPIP-based version analyzed here; the original 16PF scales were also less reliable than their IPIP counterparts (average $\alpha = .76$ vs. .83), which may have contributed to inflate the size of estimated differences. In addition, the online sample of Open Source Psychometrics was self-selected, whereas the standardization sample of the original study was designed to be demographically representative of the U.S. population (and was more balanced by sex: 50% females vs. 63%). Cohort effects may have contributed as well: the standardization sample analyzed in the original study was collected in 1993 (see Del Giudice et al., 2012), whereas the Open Source Psychometrics dataset was last updated in 2014. Finally, one has to consider the possible impact of PAV, which can both inflate and deflate the size of group differences and was not examined in the original study (e.g., in the PAV analysis of the U.S. sample, 25% of the *D* values were larger than 2.37; see Figure 1a).

The univariate effect sizes in the U.S. sample were highly correlated with those of the original study ($r = .95$ for observed scores, .90 for MG-MCSA estimates), though generally smaller in magnitude. Heterogeneity statistics indicated a somewhat more balanced contribution of individual variables to the overall effect size ($H_2 = .81$–$.83$ vs. .90 in the original study; $EPV_2 = .23$–$.24$ vs. .16 in the original study. See Del Giudice, 2018). Considering the largest univariate effects, females scored higher in Sensitivity, Anxiety (Apprehension), Warmth, and Complexity (Openness to change), whereas males were higher in Dutifulness (Rule-Consciousness), Emotional Stability, and Assertiveness (Dominance). As in

TABLE 5 Goodness-of-fit statistics for MG-CMSA models (English-speaking sample)

Model	χ^2	*df*	CFI	NNFI	RMSEA	SRMR
1. Configural invariance	81,312.73	1,680	.909	.893	.055	.055
2. Metric invariance	82,358.52	1,725	.908	.894	.054	.054
Δ 1 versus 2			*−.001*	*.001*	*−.001*	*−.001*
3. Scalar invariance	88,602.66	1,755	.901	.888	.056	.056
Δ 2 versus 3			*−.007*	*−.006*	*.002*	*.002*
4. Equality of covariances	89,637.46	1,860	.900	.893	.055	.055
Δ 3 versus 4			*−.001*	*.005*	*−.001*	*−.001*

TABLE 6 Correlations and univariate effect sizes for MG-CMSA latent scores (English-speaking sample)

	A.	C.	E.	F.	G.	H.	I.	L.	M.	N.	O.	Q1.	Q2.	Q3.	Q4.
A. Warmth		.36	.22	.51	−.29	.66	.37	−.58	−.08	−.59	−.11	−.36	−.53	−.12	−.54
C. Emot. Stability	.36		.48	.21	−.29	.55	.00	−.59	−.30	−.34	−.90	−.17	−.28	−.25	−.70
E. Assertiveness	.28	.54		.36	.11	.60	.05	−.04	.00	−.35	−.54	−.32	−.15	−.30	−.01
F. Gregariousness	.59	.30	.42		.17	.76	−.06	−.22	.20	−.45	−.17	−.18	−.67	.18	−.10
G. Dutifulness	−.29	−.29	.07	.11		−.10	.01	.41	.65	.17	.03	−.32	.16	.46	.32
H. Friendliness	.70	.57	.64	.80	−.16		.02	−.48	−.17	−.69	−.46	−.19	−.69	−.13	−.33
I. Sensitivity	.35	−.05	.05	−.10	.11	.01		−.20	.37	−.18	.07	−.77	.20	.01	−.23
L. Distrust	−.61	−.52	−.04	−.30	.44	−.48	−.14		.31	.54	.40	.12	.46	.07	.69
M. Imagination	−.11	−.38	−.05	.09	.72	−.25	.37	.38		.16	.18	−.59	.28	.44	.17
N. Reserve	−.68	−.34	−.38	−.55	.20	−.75	−.16	.55	.20		.20	.16	.56	.07	.23
O. Anxiety	−.15	−.92	−.58	−.23	.11	−.49	.06	.36	.30	.22		.19	.15	.13	.58
Q1. Complexity	−.40	−.15	−.35	−.19	−.31	−.21	−.81	.10	−.54	.21	.15		−.10	−.10	.32
Q2. Introversion	−.60	−.33	−.23	−.71	.25	−.74	.20	.54	.36	.63	.21	−.08		−.04	.25
Q3. Orderliness	−.14	−.30	−.32	.18	.53	−.16	−.02	.11	.50	.11	.20	−.07	.02		.08
Q4. Emotionality	−.55	−.68	−.05	−.14	.35	−.35	−.29	.71	.27	.30	.59	.29	.30	.15	
d	−0.36	+0.35	+0.29	+0.04	+0.33	+0.03	−0.94	+0.01	+0.06	+0.14	−0.56	−0.20	−0.04	+0.06	−0.09

Note: Females above the diagonal; males below the diagonal. Positive values of *d* indicate that males score higher than females. The letters associated with each factor are the conventional trait identifiers in the 16PF model.

the original study, the largest univariate difference was found on the Sensitivity factor (sensitive, aesthetic, sentimental, intuitive, and tender-minded vs. utilitarian, objective, unsentimental, and tough-minded).

This pattern of univariate effects is in line with previous findings on sex differences based on the Big Five model. The Sensitivity factor overlaps with both Agreeableness and "feminine openness/closedness," a composite of Openness facets that was consistently higher in women in the cross-cultural analysis by Costa and colleagues (2001). In the Big Five, Warmth is an Extraversion facet that is somewhat higher in women, whereas Assertiveness is consistently higher in men (Costa et al., 2001; Kajonius & Johnson, 2018). Sex differences in Emotional Stability map on those in (negative) Neuroticism; notably, Anxiety is the facet of Neuroticism that shows the largest sex differences (Costa et al., 2001; Kajonius & Johnson, 2018).

In both the present dataset and the original study (Del Giudice et al., 2012), males scored higher in Dutifulness, which may seem surprising since the Dutifulness facet of Big Five Conscientiousness shows higher scores in females (Costa et al., 2001; Kajonius & Johnson, 2018). However, the Dutifulness factor in the 16PF differs from the homonymous facet of the Big Five in being heavily skewed toward conservatism and respect for authority. Items of this kind in the IPIP-based version include: "I believe laws should be strictly enforced," "I resist authority" (reverse-scored), and "I like to stand during the national anthem". Accordingly, 16PF Dutifulness correlates with Big Five Conscientiousness, but also (negatively) with some Openness facets—including Actions/Adventurous and Values/Liberalism—that are typically higher in females (Conn & Rieke, 1994; Kajonius & Johnson, 2018; see also the supplementary results in Kaiser, 2019). In light of these differences, our finding of higher male scores in 16PF Dutifulness is consistent with the literature based on the Big Five.

4.1 | Implications for gender stereotypes

The present study supports the idea that global sex differences in personality are considerably larger than commonly assumed. To put our results in perspective, *D* values between 2.06 and 2.10 imply that the personality profile of a randomly picked male will be more male-typical than that of a randomly picked female about 93% of the times (common language effect size). Likewise, knowing the personality profile of an individual makes it possible to correctly guess his/her sex about 85% of the times (see Del Giudice, 2019). (Note that these figures apply to a person's "true" personality profile and not to his/her observed questionnaire scores, which are contaminated by measurement error. The corresponding probabilities for uncorrected scores are 80% and 72%.)

Of note, these findings may help answer a long-standing question in the literature (e.g., Carothers & Reis, 2013; Maney, 2016): if psychological differences are dimensional with no discrete boundaries between the sexes, why do categorical stereotypes of men's and women's behavior persist in everyday life? A possible answer is that people have a strong automatic tendency to use categorical templates to interpret the world, and for this reason misconstrue the actual structure of sex differences (Reis & Carothers, 2014). However, research on stereotypes has consistently found that people estimate sex differences in personality with high accuracy (Jussim, Crawford, & Rubinstein, 2015; Löckenhoff et al., 2014); this does not sit well with the idea that the same observers exaggerate the separation between the sexes to the point of perceiving two non-existent categories.

The existence of large multivariate differences offers an intriguing explanation of why stereotypes about male and female psychology are often categorical (or approximately so), even if the sexes overlap substantially on each individual trait. To the extent that people are paying attention to global differences (i.e., evaluating personality profiles instead of individual traits), they should *correctly* perceive a relatively sharp boundary between the sexes, with little overlap in the middle. Although categorical stereotypes remain inaccurate in a strict sense, they may provide a reasonable approximation of the degree of statistical separation between males and females in the multivariate space. To our knowledge, this hypothesis has yet to be tested in the literature on gender stereotypes. If people integrate information about personality into multivariate profiles, they should also be able to classify individuals as male or female with relatively high accuracy when given descriptions that include multiple traits. (As noted earlier, the amount of measurement error in the descriptions would limit the degree of accuracy that can be achieved in practice.) In principle, changes in classification accuracy across different combinations of traits may be exploited to make finer distinctions between alternative models of information use—for example, to determine whether people keep trait correlations into account when making inferences about a person's sex.

4.2 | Implications for theories of sex differences

Naturally, the findings of the present study do not speak directly to the biological and/or cultural origins of sex differences. Still, it is the case that researchers who emphasize the role of sociocultural factors often view sex differences as small, malleable, and overwhelmed by similarities (see Eagly & Wood, 2013; Hyde, 2014; Hyde et al., 2019). In contrast, most biologically oriented scholars argue that differences between the sexes on specific traits can be large, robust, and potentially universal (though not necessarily fixed in size), as a result of sexual selection and other evolutionary pressures that affect the sexes in divergent ways (see Archer, 2019; Buss, 1995; Schmitt, 2015).

The sociocultural malleability of sex differences is a central tenet of *social role theory* (Eagly & Wood, 1999; Wood & Eagly, 2012). The theory maintains that most sex differences in psychology and behavior arise because males and females are socialized into culturally prescribed roles, which in turn are historically based on the existence of evolved dimorphism in bodily size and function. A key prediction of social role theory is that sex differences should shrink as societies adopt more gender egalitarian values and socialization patterns. In the domain of personality, however, cross-cultural studies have generally found the opposite pattern—that is, sex differences are *magnified* in more gender egalitarian countries (Kaiser, 2019; Mac Giolla & Kajonius, 2018; Schmitt, 2015; Schmitt et al., 2016). From a biological perspective, a plausible explanation of this and similar finding (e.g., concerning values and occupational preferences) is that gender egalitarian cultures leave men and women freer to express their evolved predispositions (Schwartz & Rubel, 2005; Schmitt, 2015; Schmitt et al., 2016). At the same time, the apparent effect of increasing gender equality may be confounded with that of decreasing ecological stress in more developed countries (Kaiser, 2019). Yet another hypothesis is that, in less gender egalitarian societies, people tend to evaluate themselves using their own sex as the reference group; as gender equality increase, the reference group expands to include the entire population, thus increasing the size (and accuracy) of self-reported differences (Lippa, 2010; Lukaszewski, Roney, Mills, & Bernard, 2013; for a critical evaluation see Schmitt et al., 2016).

Considered in this context, the present results are consistent with the findings of previous cross-cultural studies based on the Big Five model. In the recent study by Kaiser (2019), MG-CMSA on 30 Big Five facets yielded $D = 2.16$ for the U.S. sample, an effect almost identical to the one we found here (effect sizes ranged from 1.49 in Pakistan to 2.48 in Russia). Likewise, the uncorrected effect size in Mac Giolla and Kajonius (2018) was $D = 1.25$ for the U.S. sample, compared to 1.18 in the present study (effect sizes ranged from 0.87 in Malaysia to 1.32 in Norway and Sweden).

4.3 | Future directions

Now that large sex differences have been found in two independent datasets based on the 16PF, it will be important to investigate the extent to which the size of the effect may depend on the choice of a personality model (e.g., 16PF vs. Big Five). So far, multivariate studies of sex differences have yielded fairly consistent results regardless of the underlying model (Kaiser, 2019; Mac Giolla & Kajonius, 2018). However, it will take a number of large-scale replications before a confident statement can be made. Moreover, to our knowledge, some popular models (e.g., the six-factor HEXACO; Lee & Ashton, 2004) have yet to be approached from a multivariate perspective. As more data accumulate, it will become possible to use meta-analysis to explore patterns of consistency and variation in a more systematic way.

Another interesting topic for future research is the impact of different modeling approaches on the estimation of latent differences. For example, exploratory structural equation modeling (ESEM; Marsh, Morin, Parker, & Kaur, 2014; Marsh et al., 2009) has been gaining popularity in recent years. Standard confirmatory models like the ones we employed in the present study constrain each item to load on one particular latent factor (the *independent clusters assumption*); in contrast, ESEM allows items to freely cross-load on all the factors. As a result, interfactor correlations tend to become substantially smaller (e.g., Booth & Hughes, 2014; Furnham, Guenole, Levine, & Chamorro-Premuzic, 2013; Marsh et al., 2010; Marsh, Nagengast, & Morin, 2013).

Given the importance of correlations among variables in the calculation of multivariate indices such as *D*, the choice of measurement model can be expected to have non-trivial consequences. Whether the ESEM approach is well suited to study sex differences in personality is not clear at this point. On the one hand, failing to model cross-loadings may lead to inflated correlations among factors (Marsh et al., 2010, 2014, 2009). On the other hand, extensive cross-loadings may end up altering the nature of the factors and "blur" their content to some extent. This is relevant because sex differences are revealed most clearly in narrow, circumscribed traits; not infrequently, traits that positively correlate with one another (e.g., different facets of Extraversion in the Big Five; Warmth and Emotional Stability in the 16PF) show sex differences of opposite sign (see Del Giudice, 2015; Del Giudice et al., 2012). If narrow traits are allowed to cross-load extensively, their specificity—and hence their ability to differentiate between males and females—may deteriorate to an unknown extent. Future research on ESEM should consider the impact of cross-loadings on both interfactor correlations and univariate differences, as well as their interplay in the determination of global sex differences.

From a theoretical standpoint, our findings corroborate those of recent multivariate cross-cultural studies, and further challenge the received view on sex differences in psychology—which, as noted, is largely modeled on the gender similarities hypothesis. Importantly, Hyde's hypothesis was framed in strictly univariate terms (Hyde, 2005, 2014); accordingly, the standard approach in the literature is to consider individual traits one at a time, or at most average them together (e.g., Zell et al., 2015). As it has become apparent over the past few years, a multivariate perspective offers a strikingly different picture of sex differences and similarities, not just in personality but in domains such as mate preferences (Conroy-Beam, Buss, Pham, & Shackelford, 2015) and occupational interests (Morris, 2016). An important research question that naturally lends itself to a multivariate approach

is the extent to which sex differences in personality predict sex differences in life outcomes such as health, well-being, and occupational choices (Soto, 2019). It is plausible that multivariate profiles will prove more predictive than individual traits, particularly if multiple aspects of personality interact in nonadditive ways to influence the relevant outcomes. In sum, we believe that the shift from an exclusively univariate focus to a multivariate one is an exciting opportunity, with the potential to dramatically improve our understanding of how personality differences play out in the lives of men and women.

ACKNOWLEDGMENTS

The authors received no financial support for the research, authorship, and/or publication of this article.

CONFLICT OF INTERESTS

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ORCID

Tim Kaiser https://orcid.org/0000-0002-6307-5347
Marco Del Giudice https://orcid.org/0000-0001-8526-1573

REFERENCES

Allen, C. N. (1930). Recent studies in sex differences. *Psychological Bulletin*, *27*, 394–407. https://doi.org/10.1037/h0070355

Archer, J. (2019). The reality and evolutionary significance of human psychological sex differences. *Biological Reviews*, *94*, 1381–1415. https://doi.org/10.1111/brv.12507

Booth, T., & Hughes, D. J. (2014). Exploratory structural equation modeling of personality data. *Assessment*, *21*, 260–271. https://doi.org/10.1177/1073191114528029

Buss, D. M. (1995). Psychological sex differences: Origins through sexual selection. *American Psychologist*, *50*, 164–171. https://doi.org/10.1037/0003-066X.50.3.164

Carothers, B. J., & Reis, H. T. (2013). Men and women are from earth: Examining the latent structure of gender. *Journal of Personality and Social Psychology*, *10*, 385–407. https://doi.org/10.1037/a0030437

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464–504. https://doi.org/10.1080/10705510701301834

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Conn, S. R., & Rieke, M. L. (Eds.). (1994). *The 16PF fifth edition technical manual*. Champagne, IL: Institute for Personality and Ability Testing Inc.

Conroy-Beam, D., Buss, D. M., Pham, M. N., & Shackelford, T. K. (2015). How sexually dimorphic are human mate preferences? *Personality and Social Psychology Bulletin*, *41*, 1082–1093. https://doi.org/10.1177/0146167215590987

Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, *81*, 322–331. https://doi.org/10.1037/0022-3514.81.2.322

Del Giudice, M. (2009). On the real magnitude of psychological sex differences. *Evolutionary Psychology*, *7*, 264–279. https://doi.org/10.1177/147470490900700209

Del Giudice, M. (2013). Multivariate misgivings: Is *D* a valid measure of group and sex differences? *Evolutionary Psychology*, *11*, 1067–1076.

Del Giudice, M. (2015). Gender differences in personality and social behavior. In J. D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (2nd ed., pp. 750–756). New York, NY: Elsevier.

Del Giudice, M. (2017). Heterogeneity coefficients for Mahalanobis' *D* as a multivariate effect size. *Multivariate Behavioral Research*, *52*, 216–221.

Del Giudice, M. (2018). Addendum to: Heterogeneity coefficients for Mahalanobis' *D* as a multivariate effect size. *Multivariate Behavioral Research*, *53*, 571–573.

Del Giudice, M. (2019). Measuring sex differences and similarities. In D. P. VanderLaan & W. I. Wong (Eds.), *Gender and sexuality development: Contemporary theory and research*. New York, NY: Springer.

Del Giudice, M., Booth, T., & Irwing, P. (2012). The distance between Mars and Venus: Measuring global sex differences in personality. *PLoS ONE*, *7*, e29265. https://doi.org/10.1371/journal.pone.0029265

Eagly, A. H., & Wood, W. (1999). The origins of sex differences in human behavior: Evolved dispositions versus social roles. *American Psychologist*, *54*, 408–423. https://doi.org/10.1037/0003-066X.54.6.408

Eagly, A. H., & Wood, W. (2013). The nature–nurture debates: 25 years of challenges in understanding the psychology of gender. *Perspectives on Psychological Science*, *8*, 340–357. https://doi.org/10.1177/1745691613484767

Feitosa, J., Joseph, D. L., & Newman, D. A. (2015). Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not English. *Personality and Individual Differences*, *75*, 47–52. https://doi.org/10.1016/j.paid.2014.11.017

Fiske, S. T. (2017). Prejudices in cultural contexts: Shared stereotypes (gender, age) versus variable stereotypes (race, ethnicity, religion). *Perspectives on Psychological Science*, *12*, 791–799. https://doi.org/10.1177/1745691617708204

Friedman, H. S., & Kern, M. L. (2014). Personality, well-being, and health. *Annual Review of Psychology*, *65*, 719–742. https://doi.org/10.1146/annurev-psych-010213-115123

Furnham, A., Guenole, N., Levine, S. Z., & Chamorro-Premuzic, T. (2013). The NEO Personality Inventory-Revised: Factor structure and gender invariance from exploratory structural equation modeling analyses in a high-stakes setting. *Assessment*, *20*, 14–23. https://doi.org/10.1177/1073191112448213

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, *7*, 7–28.

Haines, E. L., Deaux, K., & Lofaro, N. (2016). The times they are a-changing… or are they not? A comparison of gender stereotypes,

1983–2014. *Psychology of Women Quarterly*, *40*, 353–363. https://doi.org/10.1177/0361684316634081

Hess, M. R., Hogarty, K. Y., Ferron, J. M., & Kromrey, J. D. (2007). Interval estimates of multivariate effect sizes: Coverage and interval width estimates under variance heterogeneity and nonnormality. *Educational and Psychological Measurement*, *67*, 21–40. https://doi.org/10.1177/0013164406288159

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. https://doi.org/10.1080/10705519909540118

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*, 581–592. https://doi.org/10.1037/0003-066X.60.6.581

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology*, *65*, 373–398. https://doi.org/10.1146/annurev-psych-010213-115057

Hyde, J. S., Bigler, R. S., Joel, D., Tate, C. C., & van Anders, S. M. (2019). The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist*, *74*(2), 171–193. https://doi.org/10.1037/amp0000307

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2018). *semTools: Useful tools for structural equation modeling*. R package version 0.5-1. Retrieved from https://CRAN.R-project.org/package=semTools

Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (in)accuracy in perceptions of groups and individuals. *Current Directions in Psychological Science*, *24*, 490–497. https://doi.org/10.1177/0963721415605257

Kaiser, T. (2019). Nature and evoked culture: Sex differences in personality are uniquely correlated with ecological stress. *Personality and Individual Differences*, *148*, 67–72. https://doi.org/10.1016/j.paid.2019.05.011

Kajonius, P. J., & Johnson, J. (2018). Sex differences in 30 facets of the five factor model of personality in the large public (N= 320,128). *Personality and Individual Differences*, *129*, 126–130.

Kajonius, P., & Mac Giolla, E. (2017). Personality traits across countries: Support for similarities rather than differences. *PLoS ONE*, *12*, e0179646. https://doi.org/10.1371/journal.pone.0179646

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, *65*, 51–69. https://doi.org/10.1177/0013164404264850

Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin*, *136*, 768–821. https://doi.org/10.1037/a0020327

Landis, R. S., Beal, D. J., & Tesluk, P. E. (2000). A comparison of approaches to forming composite measures in structural equation models. *Organizational Behavioral Research*, *3*, 186–207. https://doi.org/10.1177/109442810032003

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, *39*, 329–358. https://doi.org/10.1207/s15327906mbr3902_8

Lippa, R. A. (2010). Gender differences in personality and interests: When, where, and why? *Social and Personality Psychology Compass*, *4*, 1098–1110. https://doi.org/10.1111/j.1751-9004.2010.00320.x

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*, 285–300. https://doi.org/10.1037/a0033266

Löckenhoff, C. E., Chan, W., McCrae, R. R., De Fruyt, F., Jussim, L., De Bolle, M., … Terracciano, A. (2014). Gender stereotypes of personality: Universal and accurate? *Journal of Cross-Cultural Psychology*, *45*, 675–694. https://doi.org/10.1177/0022022113520075

Lukaszewski, A. W., Roney, J. R., Mills, M. E., & Bernard, L. C. (2013). At the interface of social cognition and psychometrics: Manipulating the sex of the reference class modulates sex differences in personality traits. *Journal of Research in Personality*, *47*, 953–957. https://doi.org/10.1016/j.jrp.2013.08.010

Mac Giolla, E., & Kajonius, P. J. (2018). Sex differences in personality are larger in gender equal countries: Replicating and extending a surprising finding. *International Journal of Psychology*. https://doi.org/10.1002/ijop.12529

Maney, D. L. (2016). Perils and pitfalls of reporting sex differences. *Philosophical Transactions of the Royal Society of London B*, *371*, 20150119. https://doi.org/10.1098/rstb.2015.0119

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the Big-Five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*, 471–491. https://doi.org/10.1037/a0019227

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*, 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling*, *16*, 439–476. https://doi.org/10.1080/10705510903008220

Marsh, H. W., Nagengast, B., & Morin, A. J. (2013). Measurement invariance of big–five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology*, *49*, 1194–1218. https://doi.org/10.1037/a0026913

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*, 361–365. https://doi.org/10.1037/0033-2909.111.2.361

Morris, M. L. (2016). Vocational interests in the United States: Sex, age, ethnicity, and year effects. *Journal of Counseling Psychology*, *63*, 604–615. https://doi.org/10.1037/cou0000164

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*, 241–286. https://doi.org/10.1006/ceps.2000.1040

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reis, H. T., & Carothers, B. J. (2014). Black and white or shades of gray: Are gender differences categorical or dimensional? *Current Directions in Psychological Science*, *23*, 19–26. https://doi.org/10.1177/0963721413504105

Reiser, B. (2001). Confidence intervals for the Mahalanobis distance. *Communications in Statistics: Simulation and Computation*, *30*, 37–45. https://doi.org/10.1081/SAC-100001856

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of statistical software*, *48*(2), 1–36.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*, 23–74.

Schmitt, D. P. (2015). The evolution of culturally-variable sex differences: Men and women are not always different, but when they are… it appears not to result from patriarchy or sex role socialization. In T. K. Shackelford & R. D. Hansen (Eds.), *The evolution of sexuality* (pp. 221–256). Cham, Switzerland: Springer.

Schmitt, D. P., Long, A. E., McPhearson, A., O'Brien, K., Remmert, B., & Shah, S. H. (2016). Personality and gender differences in global perspective. *International Journal of Psychology*, *56*, 45–56. https://doi.org/10.1002/ijop.12265

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology*, *94*, 168–182. https://doi.org/10.1037/0022-3514.94.1.168

Schwartz, S. H., & Rubel, T. (2005). Sex differences in value priorities: Cross-cultural and multimethod studies. *Journal of Personality and Social Psychology*, *89*, 1010–1028. https://doi.org/10.1037/0022-3514.89.6.1010

Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychological Science*, *30*(5), 711–727.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology*, *100*, 330–348. https://doi.org/10.1037/a0021717

Sterba, S. K. (2019). Problems with rationales for parceling that fail to consider parcel-allocation variability. *Multivariate Behavioral Research*, *54*(2), 264–287. https://doi.org/10.1080/00273171.2018.1522497

Sterba, S. K., & MacCallum, R. C. (2010). Variability in parameter estimates and model fit across repeated allocations of items to parcels. *Multivariate Behavioral Research*, *45*, 322–358. https://doi.org/10.1080/00273171003680302

Thompson, H. B. (1903). *The mental traits of sex: An experimental investigation of the normal mind in men and women*. Chicago, IL: University of Chicago Press.

Weisberg, Y. J., DeYoung, C. G., & Hirsh, J. B. (2011). Gender differences in personality across the ten aspects of the Big Five. *Frontiers in Psychology*, *2*, 178. https://doi.org/10.3389/fpsyg.2011.00178

Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. *Advances in Experimental Social Psychology*, *46*, 55–123.

Zell, E., Krizan, Z., & Teeter, S. R. (2015). Evaluating gender similarities and differences using metasynthesis. *American Psychologist*, *70*, 10–20. https://doi.org/10.1037/a0038208

Zell, E., Strickhouser, J. E., Lane, T. N., & Teeter, S. R. (2016). Mars, Venus, or Earth? Sexism and the exaggeration of psychological gender differences. *Sex Roles*, *75*, 287–300. https://doi.org/10.1007/s11199-016-0622-1

How to cite this article: Kaiser T, Del Giudice M, Booth T. Global sex differences in personality: Replication with an open online dataset. *Journal of Personality*. 2020;88:415–429. https://doi.org/10.1111/jopy.12500